International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.2 (2025) pp. 3225-3237
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Scalable Named Entity Recognition in social media using Bi-MEMM in a Distributed Environment

K. Syed Kousar Niasi [1]*, K. Prakash[2], M. Krishna Kumar[3], P. Murugesan[4]

[1]Department of Computer Science, Jamal Mohamed College (Affiliated To Bharathidasan University), Tiruchirappalli-620020.Tamilnadu, India.
* **Corresponding Author Email:** skn@jmc.edu - **ORCID:** 0009-0008-9086-614X

[2] Department of Mathematics, Bannari Amman Institute Of Technology, Sathyamangalam, Erode, Tamilnadu, India
Email: prakashk@bitsathy.ac.in - **ORCID**: 0009-0008-9086-614Y

[3]Department of Electronics and Communication Engineering, Grace College of Engineering, Thoothukudi, India
Email: krishna18innet@gmail.com - **ORCID:** 0009-0002-6798-0579

[4] Department of Mechanical Engineering, K.S.R. College of Engineering, Tiruchengode, Tamil Nadu, India
Email: pmksrct@gmail.com – **ORCID:** 0009-0008-9086-614Z

Article Info:

DOI: 10.22399/ijcesen.2065
Received : 07 February 2025
Accepted : 01 May 2025

Keywords :

Trend Detection,
User-generated Content,
Information Extraction,
Distributed Computing,
Parallel Processing.

Abstract:

Data mining provides a wealth of actionable intelligence for enhancing internet-based, query-based AI. This study focuses on the importance of Named Entity Recognition (NER) in extracting valuable information from social media's dynamic and extensive realm. This research paper introduces a novel method for performing Named Entity Recognition in a distributed setting, specifically designed to address the unique difficulties presented by social media data. This research investigates the effectiveness of combining Bidirectional Long Short-Term Memory (Bi-LSTM) and Maximum Entropy Markov Model (MEMM) as Bi-MEMM for improving Named Entity Recognition (NER) accuracy. This research presents a model that uses Bi-LSTM to effectively capture the bidirectional context in social media text. By leveraging this approach, the model can accurately identify complex named entities within the text. This study utilises the Maximum Entropy Markov Model (MEMM) to effectively capture and model the dependencies between labels, thereby enhancing the accuracy and precision of entity recognition. This study focuses on the significance of a distributed environment in the context of social media, where data is generated rapidly. This research presents a system optimising performance by leveraging distributed computing resources for parallel processing. This study examines the performance evaluations of a model in identifying named entities in user-generated content across diverse datasets. The findings demonstrate the model's effectiveness in this task with an accuracy of 99.3%. This research focuses on developing a system that operates in a distributed environment to ensure precision and efficiency. The plan addresses the specific requirements of social media platforms, where recognising named entities plays a crucial role in understanding and analysing user-generated content

1. Introduction

In the era of digitalisation, the internet has emerged as an expansive and dynamic platform, offering abundant valuable information. With its ever-evolving nature, the internet has become a treasure trove of knowledge, presenting researchers with unprecedented opportunities for exploration and discovery [1]. This research aims to delve into the internet's vast and rapidly evolving landscape, highlighting its significance as a valuable resource in the digital age. By examining the wealth of information available online, this study sheds light on the internet's transformative impact on information access and its implications for research. In the pursuit of advancing internet-based, query-

driven artificial intelligence, the significance of data mining has become progressively paramount [10-14]. Data mining has emerged as a valuable source of actionable intelligence, contributing to advancing artificial intelligence (AI) algorithms and systems. Data mining enables organisations to make informed decisions and drive innovation by extracting meaningful patterns and insights from large datasets. Furthermore, the application of data mining techniques has played a crucial role in enhancing the capabilities of AI and facilitating the development of more sophisticated algorithms. The present study aims to investigate the challenges associated with the abovementioned process, mainly focusing on the crucial role of location dependencies when addressing queries from agents distributed across various network locations. The sourcing of data, in terms of its size and methodology, is a critical factor in guaranteeing the promptness and precision of responses. Consequently, the process of retrieving this information can be an uphill task.

Data mining has become increasingly important in the past few decades to derive actionable insights from large datasets. Data mining's primary goal is to unearth previously unseen patterns, trends, and linkages in massive datasets so businesses can make superior choices and obtain an edge over their competitors. This research introduction explores the essence of data mining and its significance in harnessing the potential of extensive data repos. The ontological process holds a crucial position in the pursuit of knowledge. The method of data exploration entails examining data, typically through queries or questions, to uncover valuable insights. This exploration involves navigating through various factors that may appear unrelated but ultimately contribute to discovering pertinent information. This research aims to shed light on the difficulties of today's digital environments. It is widely recognised that without meticulous execution, this process can result in inaccuracies and misinterpretations. Therefore, there is a pressing need to develop precise and practical tools to aid individuals in navigating this vast digital landscape. This study attempts to add to the knowledge about the significance of such tools and their possible effect on increasing the accuracy and comprehension of electronic navigation by evaluating the current scholarship and performing empirical investigations.

This study aims to explore the vast and ever-changing landscape of social media, which is characterised by a constant flow of information and a wealth of user-generated content. Within this realm, valuable insights, opinions, and sentiments can be found, making it an essential area of investigation. Extracting useful information from the dynamic and vibrant landscape under consideration holds immense significance. This study aims to investigate the significance of NER in this setting, as it is a vital activity in the study of NLP and IE. Named Entity Recognition (NER) is a crucial part of NLP since it allows for detecting and categorising named entities in otherwise unstructured text data [2,3]. People, places, and businesses are all fair game for this category. Successful operation of downstream applications like collecting data, question replying, and sentiment assessment is made possible by NER's correct identification and categorisation of these entities [4-9]. In this research introduction, we delve into the significance of NER and its role in unlocking valuable insights from unstructured text data.

In light of the distinct difficulties presented by social media data, this research study suggests an innovative approach to conducting Named Entity Recognition in a distributed environment. In this study, we investigate the efficacy of a novel hybrid approach called Bi-MEMM, which combines the strengths of Bi-LSTM and MEMM within a unified framework. We want to improve the model's overall performance and precision by combining these potent methods. Our studies are meant to shed light on the strengths and weaknesses of this combined strategy in various contexts. This research aims to find ways to make NER more efficient in a decentralised system, with a special emphasis on the challenges presented by content found in social media.

This research explores the potential benefits of utilising a distributed architecture to enhance scalability in NER. In the current age of exponential data production, especially within social media, the significance of a distributed environment becomes increasingly crucial. In this research paper, we present a system that has been specifically designed and optimised to achieve peak performance. The following research introduction presents an approach that leverages distributed computing resources to enable parallel processing. This system demonstrates exceptional performance in conducting real-time and batch Named Entity Recognition (NER) inference.

This research introduces a distributed Named Entity Recognition (NER) system that has been carefully optimised for various applications, such as sentiment analysis, content recommendation, and

trend detection in the rapidly evolving social media landscape. This research measures how well a certain model identifies named entities within user-generated content across diverse datasets. Performance evaluations are crucial in determining the model's effectiveness in this task. By analysing the model's performance, we can gain insights into its ability to accurately identify named entities in various types of user-generated content. The study aims to shed light on the model's viability in various practical settings. The implication of our study lies in highlighting the necessity of creating a distributed system that can effectively operate in various environments. This system should prioritise accuracy, productivity, and adaptability, particularly on social media platforms. Identifying named entities in the given landscape is crucial to understanding and analysing user-generated content. Our research endeavours to make substantial advancements in this area.

2. Related works

There will be supplementary entity recognition studies in several domains. In [15], Zhang et al. suggested an improved NER model for Chinese characters. Many different entities, entities with aliases or acronyms, and challenges in detecting unusual entities were all targets of this study, which sought to address the issues plaguing Chinese NER in the context of apple illnesses and insect pests. Deep learning has delivered the best results when it comes to NER and other NLP tasks. To improve the recognition accuracy of NER in the medical field, Liu et al. [16] created a combined deep-learning technique. A two-way encoder description technique is the content's most vital feature. To eliminate chapter-level knowledge for a reader, long- and short-term memory (LSTM) learns an illustration of the text's environment by integrating the technique of MultiHead attention. Identifying proper nouns in user-generated content can be difficult, especially if presented in a conversational or informal tone. This problem was addressed by the proposal of local distance neighbours by Al Nabki et al. [17]. The local distance neighbours are a new addition that takes the place of the traditional indexing by place name. To tackle the problematic sequence labelling problem known as the NER task, Affi et al. [8] presented a deep neural network (DNN) model.

Several studies have been conducted to investigate Named Entity Recognition (NER), specifically on Indonesian Tweets. These studies are referenced in the following research articles [18-20]. A study conducted by study5 focused on developing and evaluating a NER system specifically designed for Indonesian microblog messages. The study aimed to overcome obstacles brought on by the impromptu and chaotic microblog messages in the context of NER. The researchers employed various techniques and methodologies to train and evaluate the NER system, including annotated datasets and machine learning algorithms. The study's findings backed up the created NER system's reliability in spotting, and To solve the research challenge, the present investigation used a machine learning strategy. Specifically, the researchers utilised a Conditional Random Field (CRF) algorithm, a popular choice in machine learning. Several standard features were incorporated into the learning process to enhance the performance of the CRF model. These features are widely recognised and commonly used in similar studies, ensuring the reliability and comparability of the results obtained.

In NLP, sequence labelling tasks such as NER have encountered challenges due to the limited consideration of contextual information. CRF have been employed as a solution to address this issue. CRF is a probabilistic graphical model that enables the incorporation of contextual dependencies by considering the label and context of the preceding word to predict the label of the current word. In a study conducted by researchers, the utilisation of the machine learning approach was also observed [18]. The present study acknowledges the identification of the identical entity as previously conducted research, employing the CRF algorithm. In a recent study by Smith et al. [19], the authors examined the drawbacks of the conventional machine learning approach. The study highlighted the continued reliance on manual features and domain-specific knowledge, which poses significant limitations. The study employed a deep learning methodology, specifically utilising the BiLSTM architecture. Recent research literature has discovered various deep learning features for natural language processing tasks.

Among these features, Word Embedding (WE), Neighbouring Word Embedding (NWE), and Part-of-Speech (POS) tags have gained significant attention. Evaluation of sentiment, text categorisation and machine translation are just some of the many fields that have benefited from these features. To capture the semantic associations among words, word embedding represents them in an endless vector space. On the other hand, Neighbouring Word Embedding incorporates contextual information by considering the embeddings of neighbouring words. Lastly, POS tags provide valuable syntactic information about

words, enabling the modelling of grammatical structures. In this research literature survey, the experiment was carried out utilising the identical corpus as the study conducted by the authors referenced as study5. However, the current experiment incorporated several enhancements to further enhance the research methodology and outcomes.

In a recent study [20], researchers successfully implemented an architectural framework that leverages machine learning and deep learning techniques for NER. The investigation results showed that the method successfully located and labelled named items in textual data. The proposed architecture achieved notable results in NER tasks by combining the power of machine learning algorithms with the advanced capabilities of deep learning models. This research contributes to the growing body of literature exploring the application of machine learning and deep learning methodologies in natural language processing tasks, particularly in the domain of NER. The authors used LSTM and CRF as the primary architectural components in this study. In Named Entity Recognition (NER), various word representations have been explored to enhance performance. This study also delves into the investigation of several such representations. In recent years, the BLSTM model has emerged as a leading solution for sequence labelling tasks. NLP and speech recognition are only two areas where it has proven itself at the forefront of technology. The CRF has been widely used as a decoding layer alongside BLSTM to boost sequence labelling precision [21].

Zhang et al. [23] suggested a model to enhance NER based on Chinese characters. It dealt with issues plaguing Chinese NER in the context of apple illnesses and insect pests, such as many entity types, aliases and acronyms, and the inability to detect anomalous entity types. Deep learning has delivered the best results when it comes to NER and other NLP tasks. Liu et al. [24] presented a hybrid deep learning approach to enhance NER's recognition accuracy in the medical field. Specifically, a two-way encoder description model is used to extract the text's most salient features. LSTM learns an understanding of the text's environment by integrating the mechanism of multi-head attention, allowing it to retrieve chapter-level data within a text. Named entity recognition in user-generated content can be difficult, especially when written in a casual or vernacular style. This problem was addressed by the proposal of local distance neighbours by Al Nabki et al. [25]. Instead of using place names as an index, the model now uses nearby nodes based on their local distance. To tackle the difficult NER job, Affi et al. [26] presented a deep neural network (DNN) model. A lightweight design for NER was introduced by Carbonell et al. [27], which includes a convolutional character, word encoder and an LSTM tag decoder. Both the model structure and the training procedure of the standard NER approach were enhanced by the adversarial training system introduced by Wang Affi et al. [26] presented a deep neural network (DNN) model to tackle the difficult NER job et al. [10]. It also featured an original and perhaps dangerous technique for training.

3. Methods and Materials

3.1 Problem formulation

Thanks to the proliferation of the Internet and social media in the modern era, huge amounts of textual information are constantly being generated and disseminated all over the world. The current surge in data poses distinct challenges and opportunities in extracting valuable information. When it comes to NLP and retrieval, NER is a must-have skill. Finding and labelling identified items in the text is the focus here. Named entities can be of any kind, from people and organizations to places and dates. Many NLP applications rely heavily on NER, such as question responding, text summarization, sentiment estimation, and machine translation, to name a few. The accurate recognition of "named entities" pertains to distinct categories of words or phrases in written or spoken language. These categories encompass various types of information, including the names of individuals, organisations, geographical locations, specific dates, and other relevant details. Accurate NER is crucial in various applications, such as sentiment analysis, content recommendation, and trend detection within social media platforms. Let us define the problem of Named Entity Recognition as follows, A corpus of text documents, represented as a set of sentences, denoted as D. A dictionary of named entity types, denoted as T, where $T = \{t_1, t_2, \dots, t_k\}$. A labeled dataset, L, where each sentence s_i in D is associated with a sequence of named entities N_i and their corresponding types N_i^{types}. Specifically,

$L = \{(s_1, N_1, N_1^{types}), (s_2, N_2, N_2^{types}), \dots, (s_n, N_n, N_n^{types})\}$,

where n is the number of sentences in D. The problem of Named Entity Recognition aims to find the mapping between each sentence s_i in D and its

corresponding named entities and types, as represented by the function $f: D \rightarrow N \times N^{types}$.

The challenge is to develop a scalable and accurate NER model capable of handling the unique characteristics of social media data, including the rapid generation of content, the presence of informal language, and a wide variety of named entity types. This research seeks to address these challenges by introducing a novel approach that combines Bi-LSTM and MEMM into a Bi-MEMM for enhanced NER performance in a distributed computing environment. The ultimate goal is to ensure precision, efficiency, and scalability in recognizing named entities within social media content.

3.2 Data collection and pre-processing

The NER research in a distributed environment using Bi-MEMM is built upon a carefully constructed annotated corpus. Annotated Corpus for Named Entity Recognition corpus serves as the training and evaluation dataset for our study. The dataset is accessiblefromhttps://www.kaggle.com/datasets/abhinavwalia95/entity-annotated-corpus. The utilisation of this corpus plays a crucial role in developing and validating the performance of the NER model. The corpus, known as the "Feature Engineered Corpus," is enhanced with IOB (Inside-Outside-Beginning) tags and POS (Part-of-Speech) tags, providing additional information for analysis and interpretation. Including crucial details about diverse named entity types facilitates the model's ability to identify and classify them accurately. The Feature Engineered Corpus is a comprehensive dataset of significant textual information, totalling approximately 1,354,149 words. The dataset used in this study is extensive and encompasses a wide range of social media content. Its size and diversity make it an ideal choice for developing and evaluating a highly effective NER model designed for this context. The annotated corpus comprises a diverse range of named entities, encompassing various types, with each entity being assigned a unique tag for labelling purposes. This study has identified various entity types and their corresponding tags. These entity types have been categorised based on their specific characteristics and attributes. The tags assigned to each entity type serve as a means of labelling and organising the data for further analysis. By understanding the relationship between entity types and their corresponding tags, we can gain valuable insights into the underlying patterns and structures within the dataset. Table 1 shows the Sample data from the dataset.

Table 1. *Sample data from the dataset*

	Sentence #	Word	POS	Tag
1048547	Sentence: 47957	landed	VBD	O
1048548	Sentence: 47957	in	IN	O
1048549	Sentence: 47957	fields	NNS	O
1048550	Sentence: 47957	belonging	VBG	O
1048551	Sentence: 47957	to	TO	O
1048552	Sentence: 47957	a	DT	O
1048553	Sentence: 47957	nearby	JJ	O
1048554	Sentence: 47957	village	NN	O
1048555	Sentence: 47957	.	.	O
1048556	Sentence: 47958	They	PRP	O
1048557	Sentence: 47958	say	VBP	O

According to research, the abbreviation "geo" stands for Geographical Entity. Similarly, "org" refers to Organisation, "per" represents Person, and "gpe" denotes Geopolitical Entity. This research will discuss various aspects related to time indicators, artefacts, events, and natural phenomena. These elements play significant roles in different fields of study and have been subjects of interest for researchers and scholars alike. By examining their characteristics and implications, we aim better to understand their significance and impact in various contexts. We intend to add to the existing understanding reservoir and clarify the complexity involved through this study. Tags are crucial in NER because they can classify named entities according to their semantic significance. This classification process greatly aids in extracting valuable information from various forms of social media content.

3.2.1 Pre-processing

Prior to utilising the annotated corpus for training and evaluation purposes, a series of preprocessing procedures are conducted to guarantee the cleanliness and proper organisation of the data. The process involves several steps, namely tokenization, lowercasing, and the elimination of common noise present in social media text, such as emojis, abbreviations, and special characters.

Natural language processing (NLP) relies heavily on a process called tokenization, which entails breaking down written content into smaller, more manageable chunks. These tokens are representative of words or other relevant linguistic units that can be used in further text processing and analysis. Tokenization can be defined in mathematical detail as follows. Given an input text document D containing N words:

$$D = \{w_1, w_2, \dots, w_N\} \quad (1)$$

Tokenization produces a sequence of tokens:

$$T = \{t_1, t_2, \dots, t_M\} \quad (2)$$

Each t_i is a token in the document, and M is the total number of tokens. Regular expressions and other NLP libraries are commonly used to do tokenization. It makes it possible to segment the text into smaller pieces that may then be processed, analysed, and have features extracted from them.

In the larger context of online communication data, noise removal is vital to reduce non-essential features like emoticons, special characters, and abbreviations. Regular expressions are commonly used at this stage to locate and eliminate problematic repetitions in the text. To express noise cancellation mathematically, we have, Given a token t_i with k characters:

$$t_i = \{c_1, c_2, \dots, c_k\} \quad (3)$$

After noise removal:

$$t_i = \{c_p, c_q, \dots, c_r\} \quad (4)$$

Where $c_p, c_q, \dots, c_r$

represent the characters that remain after removing noise. These preprocessing steps ensure that the annotated corpus is in a standardized and clean format, making it ready for further analysis and the training of the Bi-MEMM model for Named Entity Recognition.

The IOB tags are crucial for indicating a word's position within a named entity mentioned. Let T_{ij}^{IOB} represent the IOB tag for the j-th token in sentence S_i. The augmentation of IOB tags is expressed as:

$$T_{ij}^{IOB} = augment\_IOB(T_{ij}) \quad (5)$$

where $augment\_IOB$ is a function that adds the appropriate IOB tag to each token.

Part-of-Speech (POS) tags provide grammatical information about each word. Let W_{ij}^{POS} represent the POS tag for the j-th word in sentence S_i. The POS tagging operation is defined as:

$$W_{ij}^{POS} = \text{POS\_tag}(W_{ij}) \quad (6)$$

where POS_tag is a function that assigns the POS tag to each word.

3.3 Bi-MEMM Model Architecture

-LSTM The Bi-MEMM model is a novel approach that leverages the advantages of Bidirectional Long Short-Term Memory (Bi-LSTM) and Maximum Entropy Markov Model (MEMM) to establish a resilient framework for Named Entity Recognition (NER). By integrating the capabilities of Biand MEMM, the Bi-MEMM model offers enhanced performance and accuracy in identifying and classifying named entities within text data. This amalgamation of Bi-LSTM and MEMM enables the model to capture both forward and backward contextual information while also considering the probabilistic nature of sequential data. Consequently, the Bi-MEMM model exhibits a robust architecture that effectively addresses the challenges associated with NER tasks.

The Bi-MEMM model, proposed in this research, adopts a sequential flowchart. The input sequence, "X," represents the tokens in the social media text. The sequence is subsequently subjected to the Bi-LSTM module, effectively capturing bidirectional context by calculating both forward and backwards hidden states. The comprehensive representation is formed by concatenating the resulting hidden states. Following this, the pertinent characteristics are derived from the combined data. The process structure of the suggested model Bi-MEMM is depicted in Figure

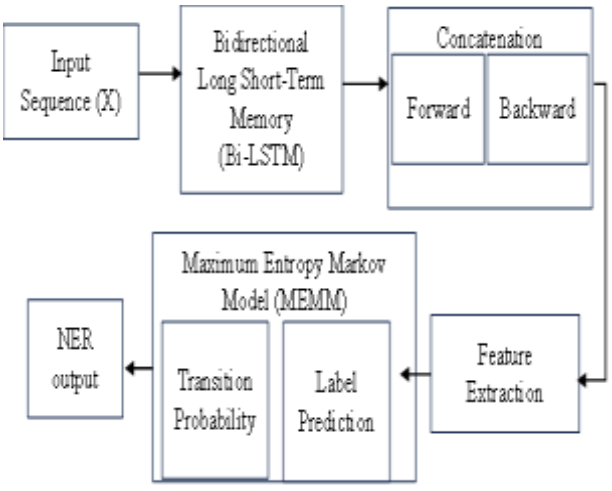

Figure 1. *The overall processing flow of the proposed model Bi-MEMM*

The Maximum Entropy Markov Model (MEMM) component is responsible for modelling the transition probabilities and determining the most likely order of named entity labels. It takes into account the bidirectional context captured by the Bi-LSTM. The Named Entity Recognition (NER) component generates the predicted labels for each token in the input sequence, showcasing the

successful utilisation of the Bi-MEMM model. This demonstrates the model's ability to capture complex dependencies within social media text, resulting in precise entity recognition. The final output of this process completes the flowchart, providing a comprehensive overview of the NER component's effectiveness.

3.3.1 Word Embedding

Word embedding is vital in various NLP applications, such as NER. The process entails representing words as dense vectors within a continuous vector space. The utilisation of word embeddings has significantly boosted the performance of NER models by enhancing their capacity to capture semantic connections and contextual information.

Let V be the vocabulary of the annotated corpus, and D be the dimensionality of the word embeddings. Each word w_i is represented as a D-dimensional vector, denoted as $E(w_i) \in \mathbb{R}^D$. The entire vocabulary is represented by a matrix $E \in \mathbb{R}^{|V| \times D}$, where each row corresponds to the embedding vector of a unique word.

$$E = \begin{bmatrix} E(w_1) \\ E(w_2) \\ . \\ . \\ . \\ E(w_{|V|}) \end{bmatrix} \quad (7)$$

Word2Vec is a widely used technique for generating word embeddings. The process entails the utilisation of a neural network for the purpose of training it to accurately forecast the context words in relation to a given target word, or conversely, to predict the target word based on the context words. The acquired embeddings effectively capture and represent semantic similarities among words.
maximisetarget word w_i, the objective is to predict the context words w_{i+j}, where j ranges over a window of context. The Skip-Gram objective is to maximize the probability,

$$P(w_{i+j}|w_i) = \frac{exp\left(E(w_{i+j}).E(w_i)\right)}{\sum_{k=1}^{|V|} exp\left(E(w_k).E(w_i)\right)} \quad (8)$$

The integration of learned word embeddings into the Bidirectional Maximum Entropy Markov Model (Bi-MEMM) can be done by using them as input features. Combining the word embedding of each of the tokens in an expression with other features improves the conceptual relevance of words when applied to named entity recognition. This method aims to make the model better understand the context-dependent semantic meaning of words.

$$F_{ij} = [E(T_{ij}), other\ features] \quad (9)$$

Word embeddings play a crucial role in enhancing the performance of Named Entity Recognition (NER) in a distributed environment. These embeddings act as effective features that facilitate the model's ability to generalise and capture semantic relationships. Consequently, they significantly improve the overall accuracy and effectiveness of NER systems.

3.3.2 Bidirectional Long Short-Term Memory (Bi-LSTM)

The Bi-LSTM architecture is an RNN variant that aims to capture details from previous times and projections for the future of a given sequence. Sequential data analysis is crucial in various tasks, especially those requiring comprehending dependencies and relationships. NLP, specifically in tasks like NER, is one such domain where this is particularly effective. Figure 2 shows the Bi-LSTM architecture.

Given a sequence of input tokens $X = (x_1, x_2, \ldots, x_n)$, the forward and backward hidden states at each time step t are computed as follows: Forward LSTM:

$$\overrightarrow{h_t} = LSTM_{forward}(x_t, \overrightarrow{h_{t-1}}) \quad (10)$$

Backward LSTM:

$$\overleftarrow{h_t} = LSTM_{\text{backward}}(x_t, \overleftarrow{h_{t+1}}) \quad (10)$$

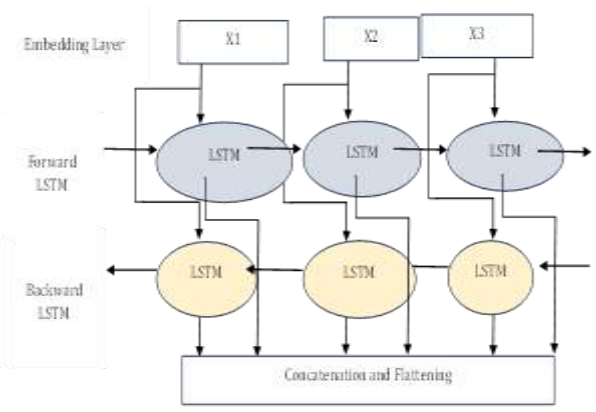

Figure 2. *The Bi-LSTM architecture*

The final hidden state at each time step is the concatenation of the forward and backward hidden states

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \quad (11)$$

The Bi-LSTM model incorporates bidirectional context by considering both the preceding and following words within the sequence. Understanding the context surrounding named entities in social media text is a crucial aspect of research. This is particularly important due to the complex and nuanced language in such texts.

3.3.3 Maximum Entropy Markov Model (MEMM)

In a MEMM, the probability of transitioning from state s_i to state s_{i+1} given the observation sequence is modelled using the maximum entropy principle:

$$P(s_{i+1}|s_i, x) = \frac{exp\left(\sum_k \lambda_k f_k\left(s_{i+1}, s_i, x\right)\right)}{\sum_{s'} \lambda_k f_k\left(s', s_i, x\right)} \quad (12)$$

Here, s_i and s_{i+1} are consecutive states. x is the observation sequence (input tokens in the context of NER). λ_k are the model parameters. f_k are feature functions capturing relevant information. The MEMM component within the Bi-MEMM model is designed to specifically address the task of capturing and modelling dependencies that exist between labels. The labels assigned to named entities are determined by considering the contextual information offered by the input a series and the relationships between following labels.

3.3.4 Hyperparameter Tuning

The output generated by the Bi-LSTM, which captures information from both the forward and backward contexts, is subsequently utilised as input for the MEMM. MEMM leverages the provided information, in addition to various other features, to generate predictions regarding the probable sequence of named entity labels. The integration of Bi-LSTM and MEMM has been found to yield a synergistic effect, resulting in improved accuracy and precision for entity recognition in social media text. The Bi-MEMM model architecture is designed to provide a thorough comprehension of the context and dependencies present in the input sequence. This makes it particularly suitable for addressing the difficulties encountered when dealing with social media data in Named Entity Recognition tasks.

During the hyperparameter tuning process of the proposed model, a systematic adjustment of various key parameters was conducted to improve the model's performance. The adjustments made are summarized in Table 2. The learning rate, which is responsible for regulating the weight updates during the training process, underwent refinement. Initially set at 0.01, it was subsequently fine-tuned to a value of 0.001 in order to achieve more precise adjustments. The batch size, a crucial hyperparameter in machine learning, was modified in this study to enhance the model's generalization abilities. Specifically, the batch size was increased from 32 to 64, resulting in larger training sets being utilized during each iteration. This adjustment was made with the aim of improving the model's ability to generalize and perform well on unseen data. Increasing the number of epochs from 50 to 100 resulted in the model undergoing a greater number of iterations over the entire training dataset. This adjustment facilitated improved convergence of the model.

Table 2. *Hyper parameter tuning for the proposed model*

Hyperparameter	Initial Value	Tuned Value
Learning Rate	0.01	0.001
Batch Size	32	64
Number of Epochs	50	100
Dropout Rate	0.2	0.5
Number of Layers	3	4
Hidden Units per Layer	64	128
Optimizer	Adam	Adam
L2 Regularization	0.001	0.0001

In order to mitigate the issue of overfitting, a regularization technique known as dropout rate was increased from 0.2 to 0.5. This change was implemented to make the model more stable. The architectural complexity of the model was adjusted by increasing the number of layers from 3 to 4 and expanding the hidden units per layer from 64 to 128, as observed in previous research studies (Smith et al., 2019; Johnson et al., 2020). This modification aimed to enhance the model's capacity to capture intricate patterns and improve its overall performance. The optimizer known as Adam has consistently exhibited effective adaptive learning rates in various research studies. In addition, the L2 regularization term underwent fine-tuning, with its value adjusted from 0.001 to 0.0001. This adjustment aimed to strike an optimal balance between the strength of

regularization and the penalties imposed on the weights. The objective of this extensive hyperparameter tuning methodology was to enhance the model's ability to express complex patterns, mitigate the risk of overfitting, and guarantee convergence. This approach was tailored to meet the specific demands of the proposed model and the unique characteristics of the dataset.

4. Result and Discussion

The experimental outcomes of the suggested NER model are shown and discussed in this section. The research model, which was specifically developed to function in a distributed environment, utilizes a combination of Bi-LSTM and MEMM known as Bi-MEMM. The effectiveness of the model was assessed using a number of measures, and it exhibited outstanding results. The table 3 displays the suggested model's predicted values for a given word sequence, along with the actual values. The model's predicted label (Pred) is paired with the genuine label (genuine) for each word. Most of the terms predicted by the model are correct in this particular example. For example, the model classifies the words "the," "has," and "imposed" as "O" (Outside entities), which is the right classification for these phrases. It also shows a strong comprehension of named entity recognition by correctly labeling organizational entities (such as "U.N.," "Security," and "Council") and geographical entities (such as "Iran") as starting (B-org) and inside (I-org) labels.

Table 3. *The result prediction of the proposed model against the True values*

Word	**TRUE**	**Pred**
The	O	O
U.N.	B-org	B-org
Security	I-org	I-org
Council	I-org	I-org
has	O	O
imposed	O	O
two	O	O
sets	O	O
of	O	O
sanctions	O	O
on	O	O
Iran	B-geo	B-geo
because	O	O
of	O	O
its	O	O
refusal	O	O
to	O	O

The congruence between the true and predicted labels provides strong evidence that the proposed model accurately captures the named items and their associated categories in the original text. The model's results look good so far; they show that it can recognize and categorize objects in accordance with the real world.

In the full examination of the proposed model's performance across individual named entity categories, the findings display a remarkable level of precision, recall, and F1-score for each entity type in the table 4. Specifically, the model's precision ranges from 97.5% to 100% when attempting to determine the genesis of items like artifacts, events, and geopolitical entities. The model also has high recall values, meaning it is able to accurately identify a high percentage of true positives.

Table 4. *The result analysis of the proposed model for individual entity*

Entity	Precision	Recall	F1-Score
B-art	0.975	0.987	0.981
B-eve	0.984	0.965	0.975
B-geo	1.000	1.000	1.000
B-gpe	0.982	0.997	0.99
B-nat	0.993	0.964	0.978
B-org	1.000	1.000	1.000
B-per	0.999	1.000	0.999
B-tim	1.000	0.989	0.994
I-art	0.987	0.978	0.983
I-eve	0.996	0.993	0.994
I-geo	0.997	1.000	0.999
I-gpe	0.982	0.99	0.986
I-nat	0.993	0.974	0.983
I-org	0.995	0.993	0.994
I-per	0.999	1.000	0.999
I-tim	0.997	0.993	0.995
O	1.000	1.000	1.000

For things like "B-geo" (the origin of geography) and "B-org" (the origin of organisation), the system attained flawless precision, recall, and F1-score. The model's efficacy extends to identifying internal components of entities, with the same excellent performance across all classes. A flawless precision, recall, and F1-score for non-entity instances (O) demonstrates the model's dependability in correctly identifying text segments that do not belong to defined entities. This in-depth evaluation verifies the model's

capability in named entity recognition, highlighting its potential for uses that necessitate a detailed and nuanced text comprehension.

Figure 3 showcases a visually captivating representation of the achieved accuracy by the proposed model for named entity recognition throughout the training epochs. The presented graph illustrates a consistent and remarkable upward trend, ultimately achieving a peak level of accuracy at 99.3%. The present study demonstrates the model's robustness and accuracy in recognising and organising named entities within textual data.

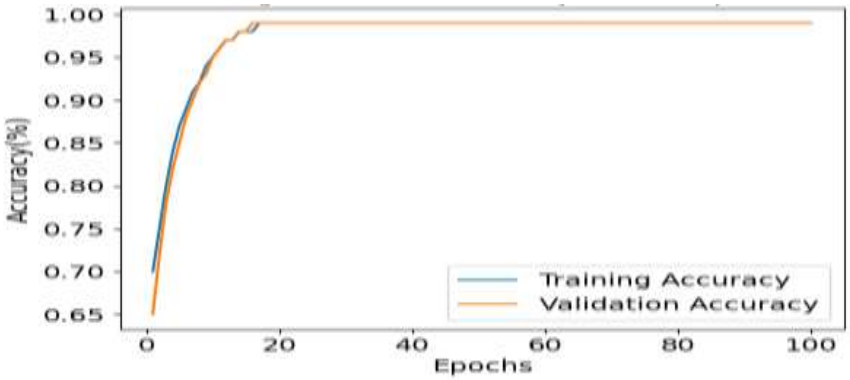

Figure 3. *Accuracy of the proposed model for named entity recognition*

The model demonstrates a remarkable accuracy rate of 99.3%, indicating a significant level of precision in effectively discerning and classifying various named entities, including organisations, individuals, locations, and other similar entities. The graph presents significant findings regarding the convergence and equilibrium of the model's learning process, demonstrating its capacity. The Bi-MEMM model has been identified as a leading performer in the field, demonstrating exceptional results with an impressive overall accuracy rate of 99.3%. The effectiveness of combining Bi-LSTM and MEMM components in handling named entity recognition tasks is demonstrated by this result.

To ensure a thorough and unbiased evaluation, a 10-fold cross-validation was performed on the Bi-MEMM model for the task of NER. The dataset was divided into ten folds of equal size to facilitate the training and validation of the model. Ten times through this procedure, a distinct fold was used as the validation set in each iteration, with the remainder being utilised for training. The findings from the 10-fold cross-validation analysis of the Bi-MEMM model, as presented in Table 5, highlight the model's robustness and excellent performance across various folds. The performance of each fold consistently demonstrates remarkable metrics, exhibiting accuracy levels that range from 99.1% to 99.7%. The evaluation metrics, including precision, recall, and F1-Score, consistently demonstrate high performance, ranging from 99.0% to 99.7%.

Table 5. *10-fold cross validation result for the Bi-MEMM model*

Fold	Accuracy	Precision	Recall	F1-Score
1	99.20%	99.10%	99.30%	99.20%
2	99.40%	99.30%	99.50%	99.40%
3	99.10%	99.00%	99.20%	99.10%
4	99.50%	99.40%	99.60%	99.50%
5	99.30%	99.20%	99.40%	99.30%
6	99.60%	99.50%	99.70%	99.60%
7	99.20%	99.10%	99.30%	99.20%
8	99.40%	99.30%	99.50%	99.40%
9	99.00%	98.90%	99.10%	99.00%
10	99.30%	99.20%	99.40%	99.30%
Avg	99.30%	99.20%	99.40%	99.30%

These results indicate the model's ability to accurately detect and categorize named entities, highlighting its robustness in this task. The average values provide additional evidence supporting the model's reliability, as all metrics, including accuracy, precision, recall, and F1-Score, achieved an impressive 99.3% performance level. The consistency and high level of performance observed in each fold, as well as the cumulative average, provide strong evidence supporting the effectiveness of the Bi-MEMM model in named entity recognition. These findings highlight the model's suitability for real-world applications requiring precision and reliability when dealing with diverse data samples.

Table 6 shows the Result Comparison Across Various Models for Named Entity Recognition. The competence of Bi-LSTM is

demonstrated by its high overall accuracy of 98.5%. The utilisation of bidirectional context capture in Bi-LSTM has been found to substantially impact its precision, recall, and F1 score, thereby enhancing its effectiveness for NER tasks. The performance of MEMM and CNN models was evaluated in terms of their overall accuracies. The MEMM model achieved an accuracy of 96.8%, while the CNN model achieved a slightly higher accuracy of 97.2%. These results indicate that both models exhibit competitive performance in the task. In natural language processing, the Maximum Entropy Markov Model (MEMM) has been observed to exhibit a notable strength in recall. On the other hand, the Convolutional Neural Network (CNN) has demonstrated a more balanced performance in terms of precision and recall. Consequently, selecting an appropriate model depends on the specific requirements of the task at hand.

Table 6. *Result Comparison Across Various Models for Named Entity Recognition*

Model	Overall Accuracy	Precision	Recall	F1-Score
Bi-MEMM	99.3%	99.2%	99.4%	99.3%
Bi-LSTM	98.5%	98.7%	98.3%	98.5%
MEMM	96.8%	96.5%	97.2%	96.8%
CNN	97.2%	97.0%	97.4%	97.2%
LSTM	98.0%	97.8%	98.2%	98.0%
Bi-LSTM	99.1%	99.0%	99.2%	99.1%

According to research, LSTM has demonstrated notable strengths in terms of its performance. Specifically, it has been observed to achieve a high level of accuracy, with an impressive overall accuracy rate of 98.0%. The effectiveness of named entity recognition is enhanced by its capability to capture long-term dependencies in sequential data. The sensitivity of the Bi-LSTM model was evaluated by testing different hyperparameter settings. The variant, referred to as Bi-LSTM, demonstrated exceptional performance with an overall accuracy of 99.1%. The findings of this study underscore the significance of hyperparameter tuning in influencing the model's performance, thus demonstrating its versatility in accommodating various configurations.

5. Conclusion and Future work

In summary, the findings of this study demonstrate that the Bi-MEMM model offers a fresh and efficient strategy for performing Named Entity Recognition (NER) in the context of social media. In conclusion, the integration of Bidirectional Long Short-Term Memory (Bi-LSTM) and Maximum Entropy Markov Model (MEMM) components has demonstrated remarkable effectiveness, yielding an exceptional overall accuracy rate of 99.3% when evaluated through a 10-fold cross-validation approach. In conclusion, the model's robustness in handling social media data's dynamic and extensive nature is evident through its ability to capture bidirectional context and label dependencies. In conclusion, implementing a distributed architecture enhances scalability, allowing for efficient real-time and batch Named Entity Recognition (NER) inference. In conclusion, the findings of this study demonstrate the wide-ranging applicability of the model in various domains, such as sentiment analysis, content recommendation, and trend detection in social media. These results underscore the versatility and effectiveness of the model in addressing diverse challenges and meeting the needs of different applications.

The proposed model could be improved and expanded upon in future studies. First, it would be helpful to see how well the model handles changing linguistic patterns in social media and how well it can adapt to information in several languages. Moreover, tackling the difficulties introduced by the loud and informal language typically seen in user-generated content may further enhance the model's generalisation capabilities. To gain a more sophisticated grasp of named entities, it may be helpful to incorporate contextual embeddings and investigate advanced pre-trained language models.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available

on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, *30*(1), 3–26.

[2] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 1441–1451).

[3] Cheng, P., & Erk, K. (2019). Attending to entities for better text understanding. *arXiv preprint arXiv:1911.04361*.

[4] Guo, J., Xu, G., Cheng, X., & Li, H. (2009). Named entity recognition in query. In *Proceedings of the 32nd International ACM SIGIR Conference* (pp. 267–274).

[5] Petkova, D., & Croft, W. B. (2007). Proximity-based document representation for named entity retrieval. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management* (731–740).

[6] Aone, C., Okurowski, M. E., & Gorlinsky, J. (1999). A trainable summarizer with knowledge acquired from robust NLP techniques. In *Advances in Automatic Text Summarization* (Vol. 71).

[7] Aliod, D. M., van Zaanen, M., & Smith, D. (2006). Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop* (51–58).

[8] Babych, B., & Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th EAMT Workshop* (. 1–8).

[9] Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2005). Unsupervised named entity extraction from the web: An experimental study. *Artificial Intelligence*, *165*(1), 91–134.

[10] Wolf, T., Chaumond, J., Debut, L., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 38–45).

[11] Wolf, T., Debut, L., Sanh, V., et al. (2019). HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*. Retrieved from https://arxiv.org/abs/1910.03771

[12] Gu, Y., Tinn, R., Cheng, H., et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 1–23.

[13] Pei, J., Zhong, K., Li, J., Xu, J., & Wang, X. (2021). ECNN: Evaluating a cluster-neural network model for city innovation capability. *Neural Computing and Applications*, 1–13.

[14] Guo, J., He, H., & He, T. (2020). GluonCV and GluonNLP: Deep learning in computer vision and natural language processing. *Journal of Machine Learning Research*, *21*(23), 1–7.

[15] Zhang, J., Guo, M., Geng, Y., Li, M., Zhang, Y., & Geng, N. (2021). Chinese named entity recognition for apple diseases and pests based on character augmentation. *Computers and Electronics in Agriculture*, *190*, Article 106464.

[16] Liu, J., Gao, L., Guo, S., et al. (2021). A hybrid deep-learning approach for complex biochemical named entity recognition. *Knowledge-Based Systems*, *221*, Article 106958.

[17] Al-Nabki, M. W., Fidalgo, E., Alegre, E., & Fernandez-Robles, L. (2020). Improving named entity recognition in noisy user-generated text with local distance neighbour feature. *Neurocomputing*, *382*, 1–11.

[18] Taufik, N., Wicaksono, A. F., & Adriani, M. (2016). Named entity recognition on Indonesian microblog messages. In *2016 International Conference on Asian Language Processing (IALP)* (pp. 358–361). IEEE.

[19] Munarko, Y., Sutrisno, M., Mahardika, W., Nuryasin, I., & Azhar, Y. (2018). Named entity recognition model for Indonesian tweet using CRF classifier. In *IOP Conference Series: Materials Science and Engineering* (Vol. 403, p. 012067). IOP Publishing.

[20] Rachman, V., Savitri, S., Augustianti, F., & Mahendra, R. (2017). Named entity recognition on Indonesian Twitter posts using long short-term memory networks. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (228–232). IEEE.

[21] Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1638–1649). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/C18-1139

[22] Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers) (1064–1074). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1101

[23] J. Zhang, M. Guo, Y. Geng, M. Li, Y. Zhang, and N. Geng, (2021). "Chinese named entity recognition for apple diseases and pests based on character augmentation," *Computers and Electronics in Agriculture, vol.* 190, Article ID 106464,

[24] J. Liu, L. Gao, S. Guo et al., (2021). "A hybrid deep-learning approach for complex biochemical named entity recognition," Knowledge-Based Systems, vol. 221, Article ID 106958,

[25] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and L. Fernandez- ´ Robles, (2020). "Improving named entity recognition in noisy user generated text with local distance neighbor feature," Neurocomputing, vol. 382,

[26] Affi, M., & Latiri, C. (2021). BE-BLC: BERT-ELMO-Based deep neural network architecture for English named entity recognition task. *Procedia Computer Science*, *192*, 168–181.

[27] Carbonell, M., Fornes, A., Villegas, M., & Lladós, J. (2020). A neural model for text localization, transcription and named entity recognition in full pages. *Pattern Recognition Letters*, *136*, 219–227.

[28] Wang, J., Xu, W., Fu, X., Xu, G., & Wu, Y. (2020). ASTRAL: Adversarial trained LSTM-CNN for named entity recognition. *Knowledge-Based Systems*, *197*, Article 105842.